Research Proposal

Research on reconfigurable digital computing-in-memory AI chips

Abstract

The rapid development of artificial intelligence technology has given rise to various applications, especially with the widespread use of deep neural networks in computer vision, speech recognition, and IoT devices. With the introduction of Transformer models, they have been widely applied in the field of natural language processing, including the popular GPT model in recent years. However, training and inference of complex neural network models come at the cost of high computational complexity, power consumption, and expenses, requiring a significant amount of GPUs, TPUs, and NPUs to support them. While these chips are optimized for specific applications, they still encounter the storage wall and power wall brought by the von Neumann architecture. To address these challenges, researchers have proposed the concept of in-memory computing architecture, integrating multiply-accumulate unit into memory to eliminate bottlenecks and significantly reduce data movement and associated power consumption. Therefore, this research proposal aims to explore the existing work on computing-in-memory AI chips and further investigate its potential applications.

Research Background

The rapid development of artificial intelligence technology has led to a growing demand for high-performance, energy-efficient computing. The traditional von Neumann architecture is struggling to meet the increasing computational requirements and evolving application scenarios [1]. For instance, in neural network training and inference, there is a need for extensive complex computations and data movements; in the field of autonomous driving, real-time processing of large volumes of images and sensor data is required; and in the Internet of Things (IoT), there is a need to handle massive amounts of sensor data. These applications place significant demands on computing performance, low latency, and energy efficiency.

In traditional computing architectures, the CPU and GPU are responsible for executing computational tasks, while the memory is used for data storage and reading. However, due to the data movement bottleneck between computation and storage [2], traditional architectures struggle to effectively utilize computational resources. Moreover, data movement accounts for a significant portion of the overall system energy consumption [4], becoming a major energy-consuming component. This leads to performance bottlenecks and power consumption. Therefore, it is necessary to explore a new computing architecture known as in-memory computing [3] to improve computing performance, reduce data movement , and meet the diverse demands of AI applications.

Currently, there are research efforts focusing on both analog computing-in-memory AI chips [5] and digital computing-in-memory AI chips [6]. For analog computing-in-memory technology, its inherent limitations in terms of accuracy and flexibility prevent it from being widely applied in various AI scenarios. On the other hand, reconfigurable digital computing-in-memory techniques provide a balance between high energy efficiency, accuracy, and flexibility required for AI chips. Therefore, this research project primarily focuses on the investigation of reconfigurable digital computing-in-memory AI chips.

Research Aim

The purpose of this research is to design a reconfigurable digital AI chip based on in-memory computing architecture to meet the demands of high-energy-efficient and compute-intensive tasks. The specific objectives include:

Improving computational performance: By integrating computing and storage within the same unit, tasks can be executed in parallel, thereby enhancing computational performance and accelerating task processing speed.

Reducing power consumption: Power consumption can be reduced by minimizing data movement and operation frequencies, as well as optimizing circuit and architecture designs.

Supporting flexibility and adaptability: Designing a reconfigurable AI

chip that supports multiple neural network models and algorithms, enabling it to adapt to different AI application scenarios and possess good flexibility and scalability.

Achieving high-bandwidth and low-latency data interaction: By tightly integrating the computing-in-memory units and designing efficient on-chip network, the aim is to reduce data movement and communication bottlenecks. This will enable high-bandwidth and low-latency data interaction, ultimately enhancing overall system performance.

This research aims to break through the limitations of traditional computing architectures and provide a higher performance, lower power consumption, and more flexible in-memory computing solution. The goal is to drive the development and application of AI technology.

Literature Review

The design of AI chips can be traced back to Eyeriss in 2016 [7]. Chen et al. optimized the energy efficiency of the entire DCNN accelerator by reconfiguring its architecture. It belongs to ASIC design and is considered the precursor to the development of AI chips in recent years. In the years following the publication of Eyeriss, there have been numerous ASIC design works, such as DNA [8], DNPU [9], UNPU [10], LNPU [11], and Evolver [12], among others. Although these ASIC design AI chips employ various methods to improve the energy efficiency of the entire accelerator, such as maximizing data reuse, exploiting the sparsity of neural networks,

quantizing neural networks, and increasing the utilization of processing element (PE) array, they still face the power consumption caused by a large amount of data movement between computation units and storage units as neural network models and computational requirements continue to grow. Therefore, during the development of ASIC design AI chips, researchers proposed the architectures of near-memory computing and in-memory computing to address this issue. Brown et al. designed NEMO-CNN, a high-performance hardware accelerator based on the near-memory computing paradigm [13]. Tu et al. proposed a reconfigurable digital computing-in-memory AI accelerator called ReDCIM to meet the requirements of high energy-efficient, high accuracy, and high flexibility for cloud AI acceleration [14]. In addition to the numerous works on digital computing-in-memory AI chips [15] [16] [17] [18], there are also many works on analog computing-in-memory AI chips, such as [19], [20], [21], [22]. Currently, due to the widespread use of Transformer models in machine translation, reading comprehension, sentiment analysis, dialogue recognition, computer vision, and other fields, achieving remarkable results, the attention mechanism of Transformer poses new challenges in terms of storage access and computation. Therefore, there are also works focusing on accelerating Transformer, such as the digital computing-in-memory Transformer accelerators TranCIM [23] and MulTCIM [26] proposed by Tu et al.

Although significant progress has been made in this field through various works, I personally believe that there is still substantial room for development and improvement in the following three aspects.

- Computing-In-Memory AI Chips support and optimize Hybrid Neural Networks.

  Most AI chips currently focus on optimizing convolutional computations [7] [8] [12] [24] [25], and there is still significant room for improvement in simultaneously optimizing computing-in-memory AI chips for convolutional neural networks, fully connected neural networks, recurrent neural networks and so on. This is because different neural networks have different computational requirements [9]. For convolutional layers, they require a large number of computations and relatively fewer filter weights, while fully connected layers require relatively fewer computations and a large number of filter weights.

- Computing-In-Memory AI chips accelerate the Transformer model.

  The attention mechanism of the Transformer introduces dynamic matrix multiplication, where the weights and inputs are generated at runtime, resulting in off-chip storage access for intermediate data. In contrast, traditional neural networks employ static matrix multiplication, where the weights are pre-trained and remain

unchanged during the inference process. Therefore, due to the various new challenges brought by the Transformer model, research on computing-in-memory AI chips for accelerating Transformers is a hot topic for the foreseeable future.

- Computing-in-memory AI chips that support neural network training.

  Most of the currently proposed computing-in-memory AI chips have only optimized the inference of deep neural networks and do not support neural network training. As neural network training requires significant computations and memory bandwidth, comparable to neural network inference, achieving and supporting neural network training on computing-in-memory chips has the potential to greatly improve energy efficiency. There are ongoing efforts focusing on the research of computing-in-memory AI chips that support neural network training. For example, Jiang et al. proposed a computing-in-memory architecture called CIMAT for DNN training [27]. This direction still has significant room for optimization and development in the future.

Research Methods

Designing a computing-in-memory AI chip involves considering several issues and steps, including but not limited to the following:

1. Literature review: Read relevant published literatures to understand the latest research advancements, technical challenges, and future directions in the field.
2. Determination of design objectives and requirements: Based on the research objectives and application scenarios, clearly define the design goals and requirements of the computing-in-memory AI chip. This involves specifying performance metrics (such as latency, throughput), power constraints, and resource utilization requirements.
3. Hardware architecture design: Based on the defined design goals and requirements, design the hardware architecture of the computing-in-memory AI chip. This involves determining the organization and implementation of the computing-in-memory units, as well as architecture-level optimization strategies.
4. Performance evaluation and optimization: Perform pre-tapeout performance evaluation to validate the performance and effectiveness of the designed computing-in-memory AI chip. This can be done using simulators developed from existing works and measuring key performance metrics. Based on the experimental results, optimize the design, improve the hardware architecture to enhance performance, and meet the design goals.
5. Application scenario validation: After receiving the taped-out AI

chip, test and evaluate the performance and effectiveness of the computing-in-memory AI chip in real-world application scenarios. This should include the application scenarios defined during the chip design process. Perform comparative analysis of the performance differences between the computing-in-memory AI chip and traditional architectures or previous related works. Assess its advantages and potential in practical applications.

Regarding research on AI chips, I also have some personal views on the research approach. Firstly, computing-in-memory AI chips require cross-level collaborative optimization, involving device, circuit, architecture, and algorithm innovations. Optimizing a single aspect alone is difficult to achieve high energy efficiency. Secondly, the development of computing-in-memory chips needs to be combined with emerging technologies, such as computing-in-memory AI chips based on Chiplet technology.

Expected Results

Designing innovative and reconfigurable digital computing-in-memory AI chips one after another, meeting the requirements of high energy efficiency, high accuracy, and high flexibility, will drive the development and application of AI technology. It will also provide a foundation and reference for the research and development of future intelligent hardware. Of course, to facilitate developers in utilizing

computing-in-memory AI chips for various AI applications, a comprehensive software development framework and tools need to be provided. However, this is a concern for the industry, and in academia, it may be feasible to implement a simple software development framework for basic demonstrations.

Each of the above small goals requires the joint efforts of the supervisor and all members of the research group. Let's achieve it together, come on!

References

[1] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436-444.

[2] Sze, Vivienne, et al. "Efficient processing of deep neural networks: A tutorial and survey." Proceedings of the IEEE 105.12 (2017): 2295-2329.

[3] Ielmini, Daniele, and H-S. Philip Wong. "In-memory computing with resistive switching devices." Nature electronics 1.6 (2018): 333-343.

[4] Boroumand, Amirali, et al. "Google workloads for consumer devices: Mitigating data movement bottlenecks." Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems. 2018.

[5] Chen, Wei-Hao, et al. "A 65nm 1Mb nonvolatile computing-in-memory ReRAM macro with sub-16ns multiply-and-accumulate for binary DNN AI edge processors." 2018 IEEE International Solid-State Circuits

Conference-(ISSCC). IEEE, 2018.

[6] Tu, Fengbin, et al. "A 28nm 29.2 TFLOPS/W BF16 and 36.5 TOPS/W INT8 reconfigurable digital CIM processor with unified FP/INT pipeline and bitwise in-memory booth multiplication for cloud deep learning acceleration." 2022 IEEE International Solid-State Circuits Conference (ISSCC). Vol. 65. IEEE, 2022.

[7] Chen, Yu-Hsin, et al. "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks." IEEE journal of solid-state circuits 52.1 (2016): 127-138.

[8] Tu, Fengbin, et al. "Deep convolutional neural network architecture with reconfigurable computation patterns." IEEE Transactions on Very Large Scale Integration (VLSI) Systems 25.8 (2017): 2220-2233.

[9] Shin, Dongjoo, et al. "14.2 DNPU: An 8.1 TOPS/W reconfigurable CNN-RNN processor for general-purpose deep neural networks." 2017 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2017.

[10] Lee, Jinmook, et al. "UNPU: A 50.6 TOPS/W unified deep neural network accelerator with 1b-to-16b fully-variable weight bit-precision." 2018 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2018.

[11] Lee, Jinsu, et al. "7.7 LNPU: A 25.3 TFLOPS/W sparse deep-neural-network learning processor with fine-grained mixed precision of FP8-FP16." 2019 IEEE International Solid-State Circuits Conference-(ISSCC).

IEEE, 2019.

[12] Tu, Fengbin, et al. "Evolver: A deep learning processor with on-device quantization–voltage–frequency tuning." IEEE Journal of Solid-State Circuits 56.2 (2020): 658-673.

[13] Brown, Grant, Valerio Tenace, and Pierre-Emmanuel Gaillardon. "NEMO-CNN: An efficient near-memory accelerator for convolutional neural networks." 2021 IEEE 32nd International Conference on Application-specific Systems, Architectures and Processors (ASAP). IEEE, 2021.

[14] Tu, Fengbin, et al. "ReDCIM: Reconfigurable Digital Computing-In-Memory Processor With Unified FP/INT Pipeline for Cloud AI Acceleration." IEEE Journal of Solid-State Circuits 58.1 (2022): 243-255.

[15] Yin, Shihui, et al. "XNOR-SRAM: In-memory computing SRAM macro for binary/ternary deep neural networks." IEEE Journal of Solid-State Circuits 55.6 (2020): 1733-1743.

[16] Si, Xin, et al. "15.5 A 28nm 64Kb 6T SRAM computing-in-memory macro with 8b MAC operation for AI edge chips." 2020 IEEE international solid-state circuits conference-(ISSCC). IEEE, 2020.

[17] Su, Jian-Wei, et al. "Two-way transpose multibit 6T SRAM computing-in-memory macro for inference-training AI edge chips." IEEE Journal of Solid-State Circuits 57.2 (2021): 609-624.

[18] Chih, Yu-Der, et al. "16.4 An 89TOPS/W and 16.3 TOPS/mm 2 all-

digital SRAM-based full-precision compute-in memory macro in 22nm for machine-learning edge applications." 2021 IEEE International Solid-State Circuits Conference (ISSCC). Vol. 64. IEEE, 2021.

[19] Chen, Wei-Hao, et al. "A 65nm 1Mb nonvolatile computing-in-memory ReRAM macro with sub-16ns multiply-and-accumulate for binary DNN AI edge processors." 2018 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2018.

[20] Xue, Cheng-Xin, et al. "24.1 A 1Mb multibit ReRAM computing-in-memory macro with 14.6 ns parallel MAC computing time for CNN based AI edge processors." 2019 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2019.

[21] Xue, Cheng-Xin, et al. "16.1 A 22nm 4Mb 8b-precision ReRAM computing-in-memory macro with 11.91 to 195.7 TOPS/W for tiny AI edge devices." 2021 IEEE International Solid-State Circuits Conference (ISSCC). Vol. 64. IEEE, 2021.

[22] Liu, Qi, et al. "33.2 A fully integrated analog ReRAM based 78.4 TOPS/W compute-in-memory chip with fully parallel MAC computing." 2020 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2020.

[23] Tu, Fengbin, et al. "TranCIM: Full-Digital Bitline-Transpose CIM-based Sparse Transformer Accelerator With Pipeline/Parallel Reconfigurable Modes." IEEE Journal of Solid-State Circuits (2022).

[24] Chen, Yu-Hsin, et al. "Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices." IEEE Journal on Emerging and Selected Topics in Circuits and Systems 9.2 (2019): 292-308.

[25] Moons, Bert, et al. "14.5 envision: A 0.26-to-10tops/w subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28nm fdsoi." 2017 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2017.

[26] Tu, Fengbin, et al. "16.1 MuITCIM: A 28nm $2.24\mu\mathrm {J} $/Token Attention-Token-Bit Hybrid Sparse Digital CIM-Based Accelerator for Multimodal Transformers." 2023 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2023.

[27] Jiang, Hongwu, et al. "CIMAT: A compute-in-memory architecture for on-chip training based on transpose SRAM arrays." IEEE Transactions on Computers 69.7 (2020): 944-954.